IntelComp Tools

Jesús Cid-Sueiro

24/11/2022

AI living labs.

MOTIVATION AND VISION

Tools for STI analysis

- At the core of the **IntelComp** platform, there are ***tools*** for the ***processing and analysis*** of ***large document collections***
- These collections are associated to the ***STI activity*** of a group or researchers, an institution, a region, a country, a public administration…
  - papers
  - patents
  - research projects
  - clinical studies
  - code repositories,
  - company websites
  - funding agency work programmes
  - …

Statistical analysis

- **IntelComp** incorporates tools based on the standard ***statistical analysis*** of ***metadata*** information available in many open repositories (authors, affiliations, funding institutions, publication dates, keyword) and also in the word frequencies in text components (titles, abstracts, etc).
- However, classical statistical analysis has some major limitations:

Motivation: some limitations of statistical approaches

- To ***classify*** documents using metadata:
  - Taxonomy updates is (normally) a slow process, so emerging technologies may not be well covered by them
  - Different datasets use heterogeneous taxonomies (IPC-AJSC-Cordis classification-NSF/NIH class. …), hindering a joint analysis
  - Labels may lack from enough granularity.
  - Problems with classification time evolution: speed of adaptation, backward compatibility (ex. IPC versions).

Motivation: some limitations of statistical approaches

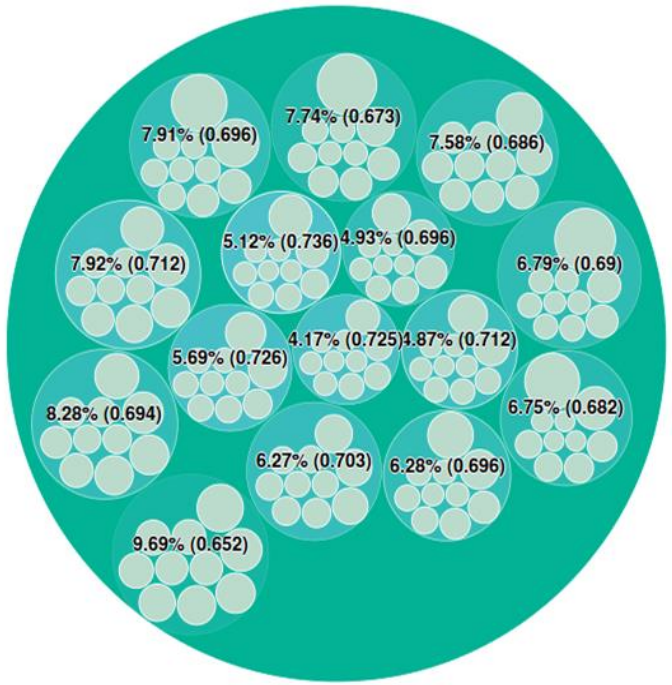
- To ***connect*** the components (documents, agents, entities) of the STI ecosystem
  - Binary document classifications are not useful to determine semantic similarities across documents
  - Poor document representation; binary multicomponent vector of classification labels (with few human defined components)

Motivation: the IntelComp vision

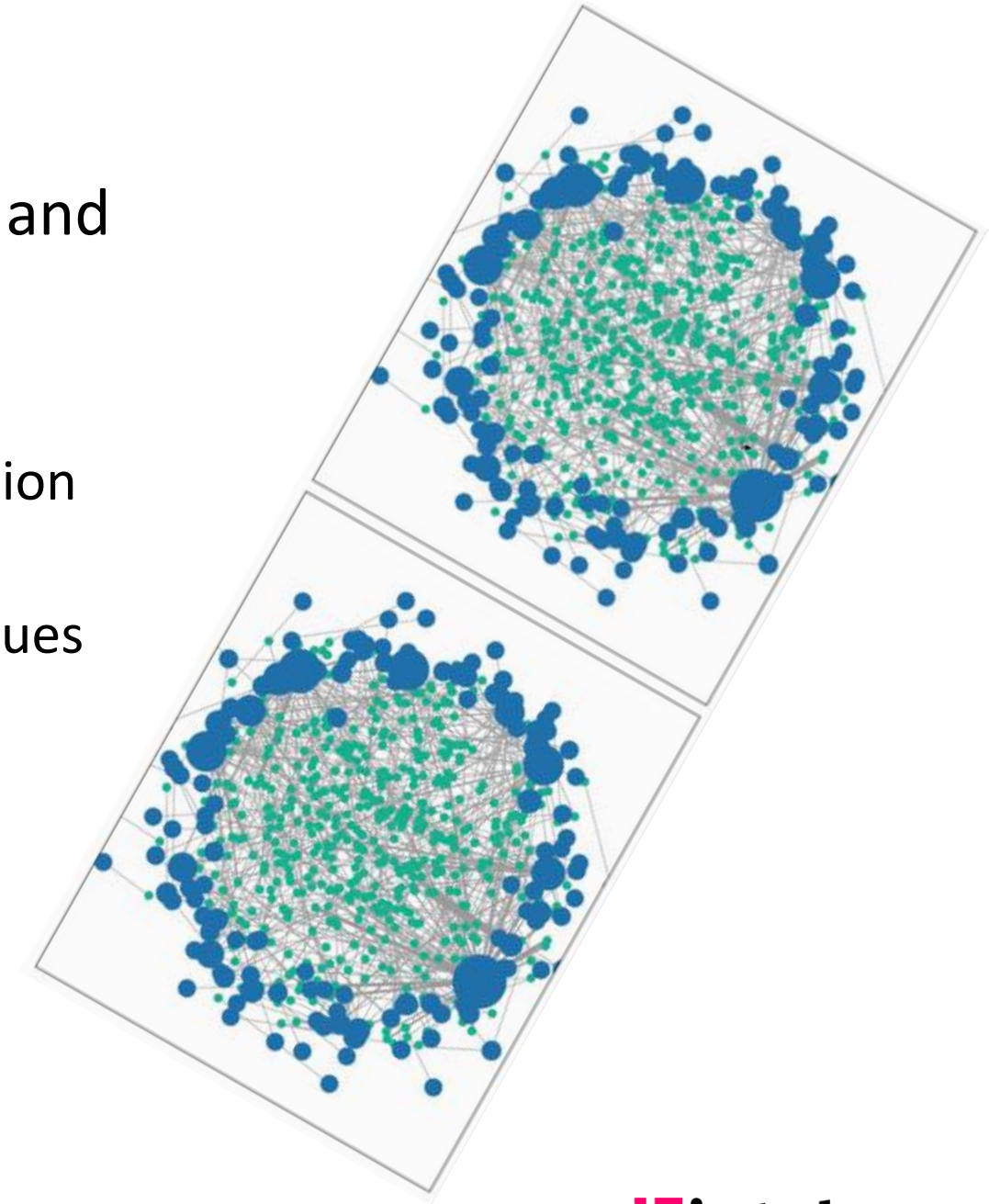
- Advanced technologies for
  - Natural Language Processing
  - Deep learning and Artificial Intelligence
  - Graph analysis and processing
  - Highly Parallel Computing
- can be used to exploit information from raw text sources and obtain much more meaningful:
  - Semantic document representations
  - Categorization of documents beyond standard taxonomies
  - Semantic connections between documents from heterogeneous sources.
  - …

NLP and AI technologies

- Scalable multilingual **NLP pipelines**
- Last-generation neural-based language models for:
  - **Domain classification**
  - **Topic modelling** (static, dynamic, hierarchical)
  - **Automatic classification**
- Information retrieval based on topic similarity and keyword search



- Scalable document **graph** generation and analysis (GPU implementations)
- **Agent analysis**:
  - Agent disambiguation and characterisation (authors, organisations)
  - Agent profiling with embedding techniques
- **Temporal analysis**
  - Lead-lag detection (thematic lead / lag between corpora)
  - Short- and Long-term Impact analysis



IntelComp is, thus,

a **Platform** for **Public Administration** (and other STI agents),

that provides **tools** for evidence-based **STI policy** (at all phases)

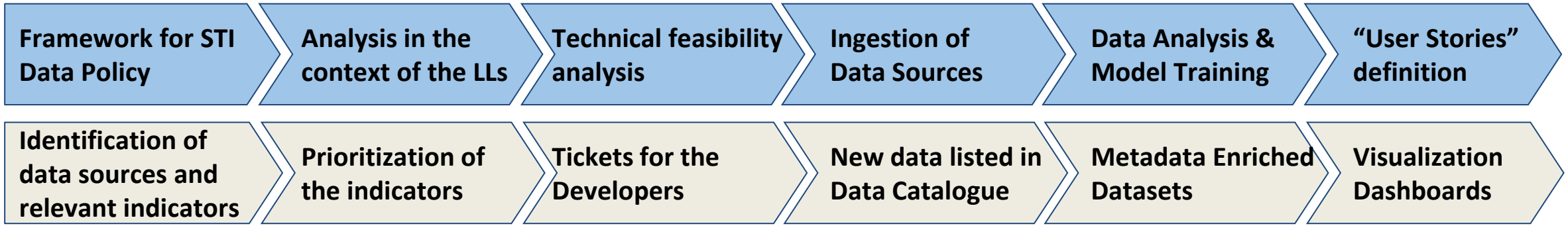
which is based on **Open data**

and is supported by **Innovative analytics services**,

NLP pipelines and AI workflows, and

deployment in **HPC and cloud** infrastructure

IntelComp's Workflow



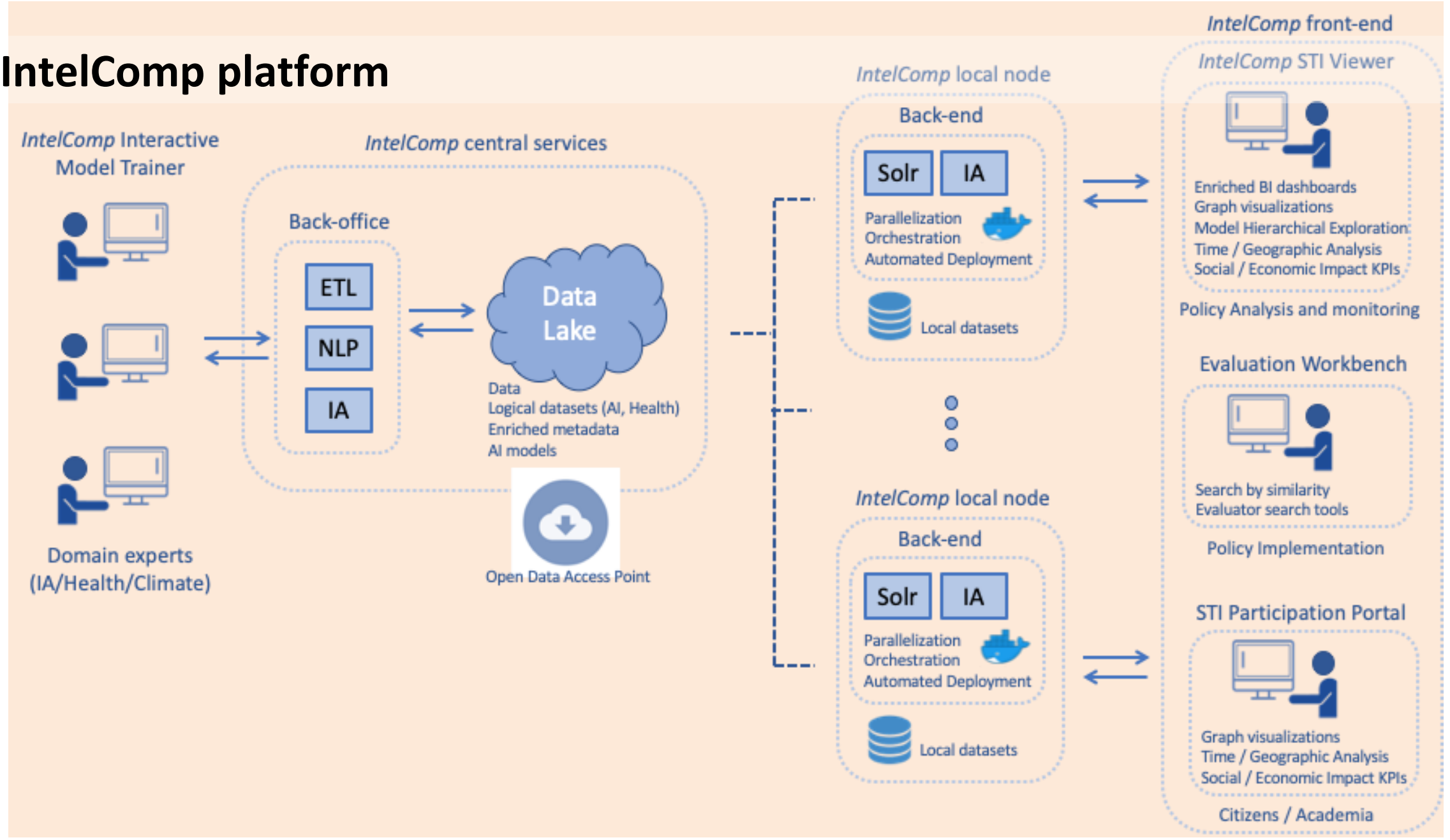
Policy Qs: Agenda Setting: Intelligence gathering, problem identification

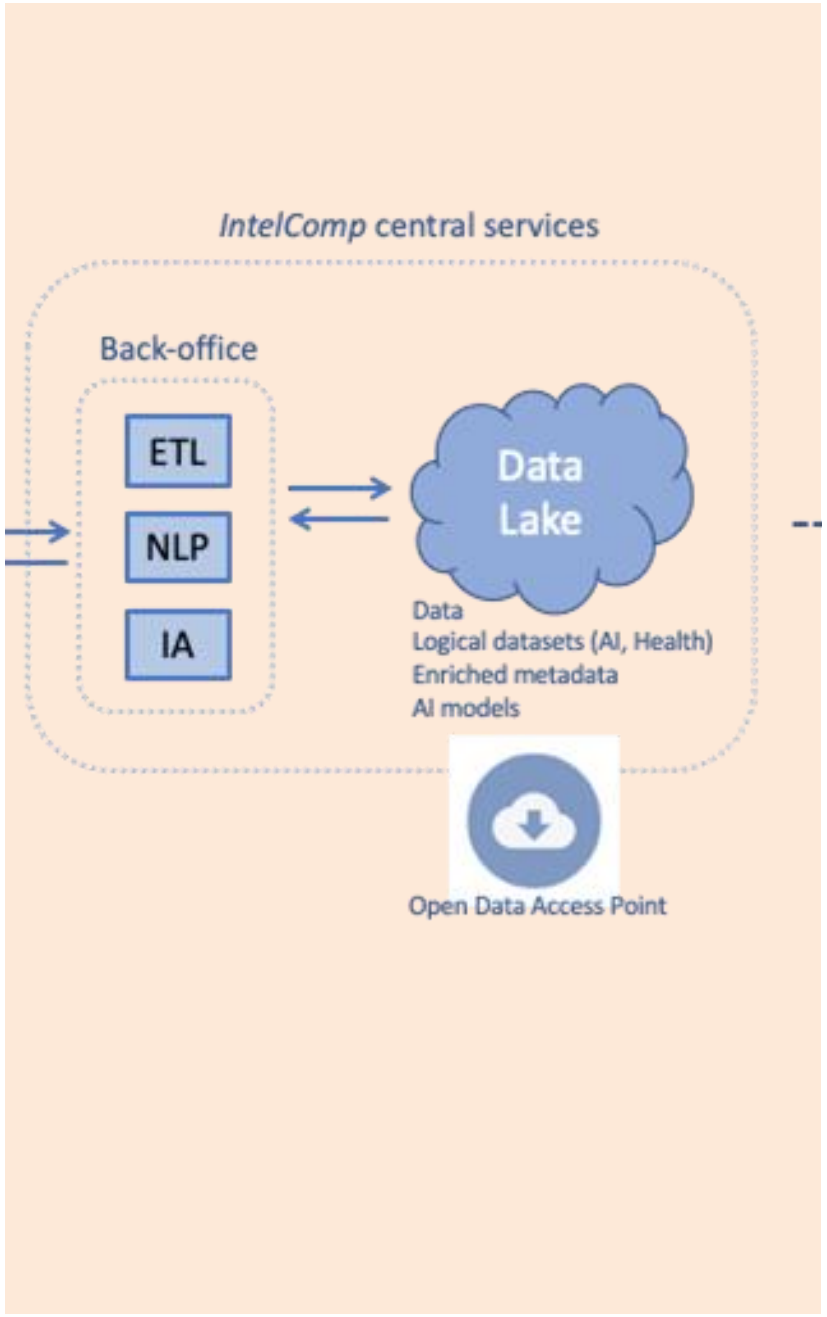
Specific Policy Questions can be addressed by matching datasets and AI-services

Area	Policy Question	Data Sets	IntelComp Services
Entrepre-neurial Activity	• **Are companies adapting to technological transformation trends in their respective sectors?** • How do they compare with major (int.) competitors? • Which companies emerge with specific disruptive technologies in the region?	**Company Websites, Scientific papers, Patents**	• **Detect technological transformation trends in a specific sector (i.e., AI techniques or AI applications)** • **Temporal evolution of topics** • Detect lead-lags in comparison of technologies across companies • Comparison of topic distribution in corpus by region.
Knowledge Creation	• Which scientific fields demonstrate the highest growth in terms of publications/citations globally? • Which are the emerging interdisciplinary fields globally? • What teams undertake research in these fields?	Scientific papers Grants (Nat., EU, Int) Patents	• Distribution of topics in corpus • Temporal evolution of topics (detection of "new" topics) • Impact analysis based on documents and authors (citations, topics)
Guidance	• To which global, EU societal challenges (i.e. living lab specific) are research groups contributing to?	Scientific papers Grants	• Specific EU societal challenge identification in corpus (i.e. public grants) • Classification of public funding proposals & grants by EU societal challenge based on a challenge description
Market	• What is the role of public procurement for these technologies (theoretically/practically)?	EC / National Public tenders	• Identification of AI techniques and AI applications in corpus
Resource mobiliz-ation	• **What financial resources are available in country?** • Are they used to leverage EU funding through synergies? • Is there a gap between supply and demand?	**EU grants** **National grants** Job postings	• **Presence of AI techniques & apps in EU / nat. grants** • Comparison of public funding in corpus (i.e. EU / national grants) by topic in a specific sector. • Comparison of topic distribution (job supply vs demand).

THE INTELCOMP PLATFORM

The IntelComp platform



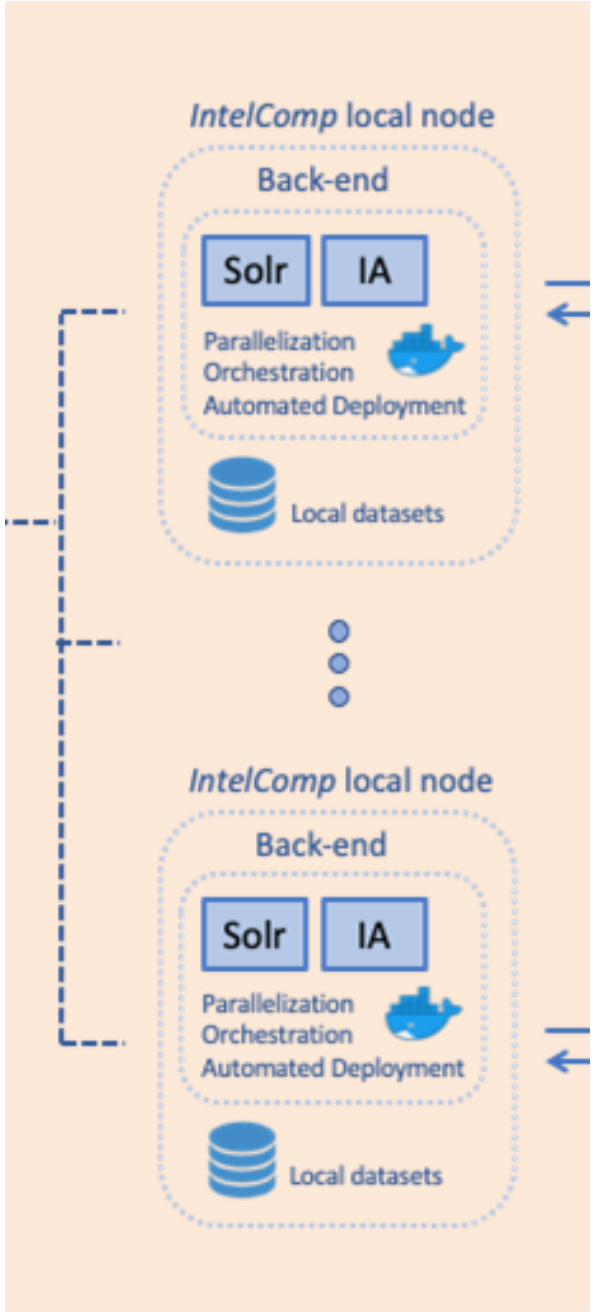


IntelComp central services

- Responsible for **highly demanding computation models**
- **Open data**: projects, articles, patents, company information, job offers, social media
- Categorized by common taxonomies and inferred topics using AI
- Possibility to create logical data sets and domains (e.g., country, dates)

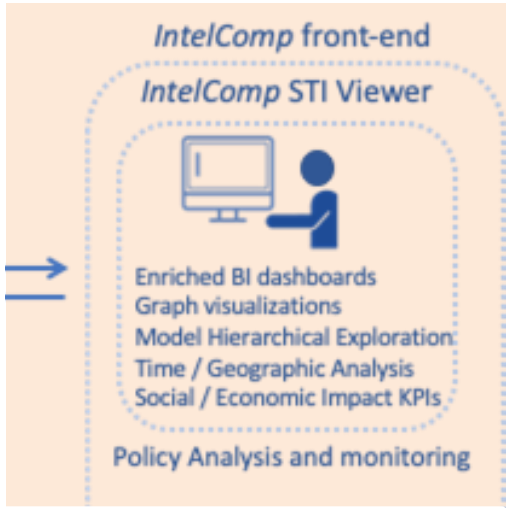
Federated local nodes

- will deploy the backend for supporting the operation of different web services.
- can coordinate by sharing only the public information from the local datasets, or by calculating intermediate data representations to ensure privacy.



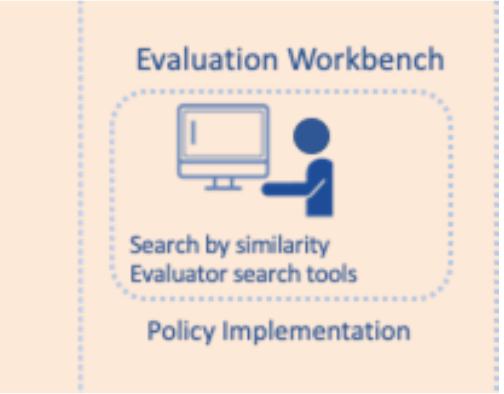
STI Viewer

- **Main tool** for **Public Policy Analysis, Planning and Monitoring**
- It assists STI policy analysts in the agenda setting and monitoring stages of the policy cycle, by mapping S&T fields, and by linking funding to outputs, outcomes and impacts.
- **Visualization tools co-created with the living labs** (agile methodology, user in the loop)
  - Topic and graph enriched BI dashboards
  - Time and geographic STI analysis
  - Comparison of different corpus (e.g., EU vs USA)
  - Profiling of STI agents: researchers, organizations
  - Analysis of projects by instrument, management unit, etc
  - Exportation of results



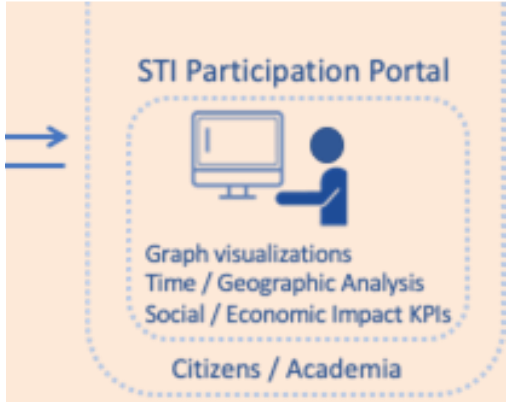
Evaluation Workbench:

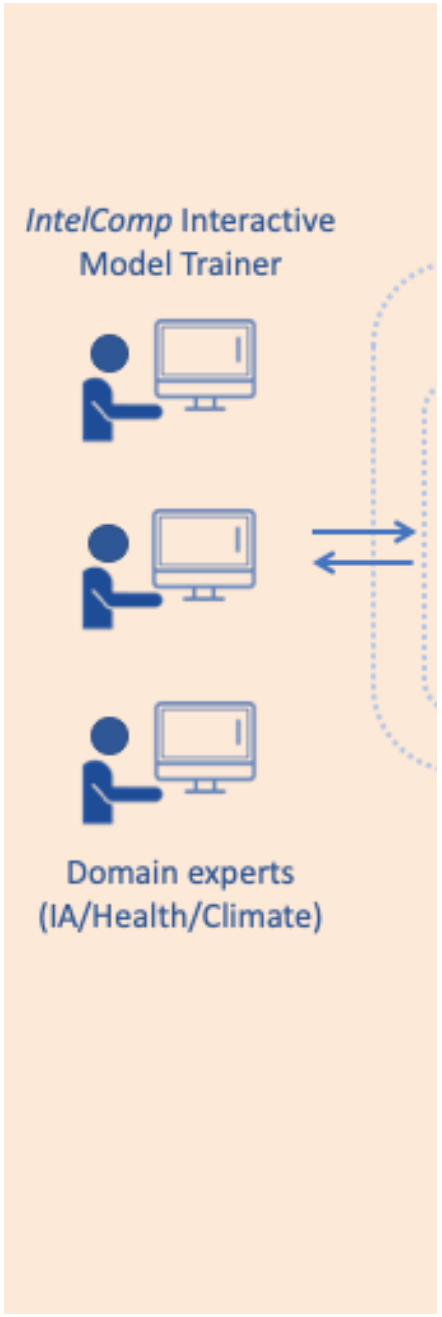
- Set of tools and functionalities for the **evaluation of grant proposals**, intended for policy implementation
- It assists call for funding managers in analyzing the SoA related to a given proposal for funding, or in detecting similar proposals or grants.



STI Participation Portal:

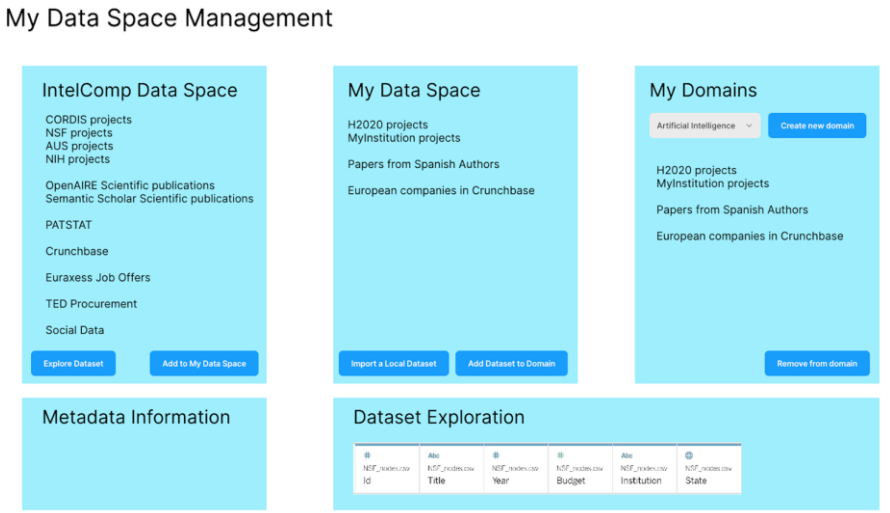
- Service targeted to two different user profiles:
  - 1) **Transparency / participation portal** for users, and
  - 2) **STI prospection** for Academia and industry.
- It assists all relevant stakeholders of STI policy (academia, industry and citizens) to co-create STI policies, synthesizing public inputs and connecting S&T fields with social needs.

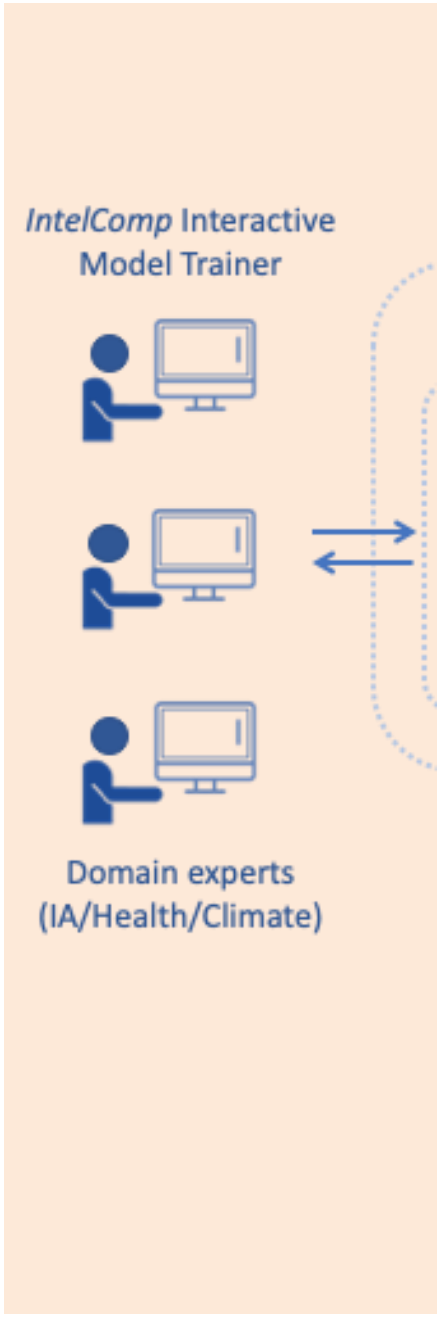




Interactive Model Trainer

- It is the ***back-office tool*** that implements curation and annotation services and intended to facilitate the construction of high-quality models by **domain experts**
- It helps academic experts in specific domains (e. g. AI, blue economy, cancer) to validate the results of the different services, to optimize them, and to design the datasets and text analysis services feeding the other three IntelComp tools





Interactive Model Trainer

- It includes
  - Definition of **target datasets**
  - **Selection** of documents by target domain
  - **Training** of models
  - Tools for **Curation & Annotation** of Models
  - **Application** of available models
  - Configuration of predefined views
- Living labs will deliver open models for the three target domains: AI, Cancer, Climate Change

THE INTELCOMP CONTEXT – END USER TOOLS

	STI Viewer	STI Policy Participation Portal	Evaluation Workbench
Targeted Organization	Public administration (Ministry), funding agency	Ministry, funding agency, academic, business and citizen organizations	Funding Agency
Targeted user	Policy & STI analyst	Policy officer, STI managers/agents for organizations, citizens	Call Manager
Main functionality	Analyze, compare and visualize a comprehensive set of STI related KPIs	To provide a synthetic list of measurements for participatory STI policy making	To assist in the ex-ante evaluation of STI proposals for funding
Stage of the policy-making cycle	Agenda setting, monitoring, evaluation	Agenda setting, monitoring, evaluation	Implementation
Previous Tool	Data4Impact	*(simplified)* STI Viewer	Corpus Viewer

intelcomp

https://intelcomp.eu/, @IntelComp_2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004870. H2020-SC6-GOVERNANCE-2018-2019-2020 / H2020-SC6-GOVERNANCE-2020